Readme for Munson et al. PNAS 2016

All sequences were generated on the Illumina MiSeq using V2 chemistry and 2x250 read lengths.

PCR amplicon structure and composition:

All samples were generated as PCR amplicons with a structure/orientation seen below:

```
Bridge---INDEXX---R2---alphaTCR---overlap---betaTCR---barcode---R1---Bridge
                          C---J---V                 V---J---C
```

Bridge – Illumina Bridge sites, V2 chemistry

INDEXX – 6 base indexes used for demultiplexing by the instrument during sequencing. The 5'->3' sequences are below:

- CGTGAT
- ACATCG
- GCCTAA
- TGGTCA
- GATCGC
- ATCATC

R1 and R2 - the Illumina read priming sites, V2 chemistry

alphaTCR – primers were designed to anneal to the C region as close to the C/J junction as possible. The orientation of the TCR is indicated above in relation to the read priming site.

overlap – sequence that was engineered to allow for overlap extension during RT-PCR. This sequence was also used in the post processing stage. 5'->3' refers to the sequence in the above structure/orientation.

- 5' - AATCAGGGACAACCTGCCCAAT – 3'

betaTCR – primers were designed to anneal to the C region as close to the C/J junction as possible. The orientation of the TCR is indicated above in relation to the read priming site.

barcode – this sequence was inserted after the R1 priming site for two reasons. Firstly, various number of randomized nucleotides were added to provide cluster diversity (as this is amplicon sequencing). Secondly, 4 base barcodes were inserted to further multiplex each Illumina barcode.

- 5’ – GAGG – 3’
- 5’ – ACGTNNNN – 3’
- 5’ – TGCANNNNNN – 3’
- 5’ – GTACNNNNNNNN – 3’

The sequences above are oriented according to the structure/orientation above, where the C region priming site would be on the 5’ side and the R1 priming site would be on the 3’ side.

Data processing strategy:

Paired end data were first parsed into multiple files based on the Illumina indicies used on the MiSeq instrument.

Read 1 and 2 sequences were clipped using a portion of the overlap sequence. AATCAGGGACAA was used for the Read 2 (alpha) sequences whereas ATTGGGCAGG was used for the Read 1 (beta) sequence file.

After the sequences were clipped, only TCR sequence remained from the TCR of interest. This step was needed given the overall size of the amplicon (450 ± 50 bp), read lengths of 250 bp would read through a TCR and into the reverse complement of the paired TCR. During our initial analysis, it became apparent that this excess sequence confounded calling of the correct TCR by our analysis pipeline.

The Read 1 (beta) files were then further demultiplexed using the barcode sequences indicated above. To do so, the GAGG barcode was first selected out. The original beta file was then shortened by 4 bases and the ACGT barcode was selected. The original file was then trimmed by 6 bases and TGCA barcode was selected. Lastly, the original file was trimmed by 8 bases and the GTAC barcode was used. All barcode selection and trimming was done from the beginning of the sequence file.

To generate the paired files, we used the MiTCR derived program CompleteTCR to pair alpha and beta TCR sequences using the cluster IDs generated during paired end sequencing. For more information see the manuscript as well as the original MiTCR publication (Bolotin DA*, et al.* (2013) MiTCR: software for T-cell receptor sequencing data analysis. *Nat Methods* 10(9):813-814).